用户画像的含义

用户画像（ persona ）的概念最早由交互设计之父 Alan Cooper 提出：“ Personas are a concrete representation of target users. 是指真实用户的虚拟代表，是建立在一系列属性数据之上的目标用户模型。随着互联网的发展，现在我们说的用户画像又包含了新的内涵 —— 通常用户画像是根据用户人口学特征、网络浏览内容、网络社交活动和消费行为等信息而抽象出的一个标签化的用户模型。构建用户画像的核心工作，主要是利用存储在服务器上的海量日志和数据库里的大量数据进行分析和挖掘，给用户贴 “标签 ”，而 “标签 ”是能表示用户某一维度特征的标识。具体的标签形式可以参考下图某网站给其中一个用户打的标签。

用户画像的作用

提取用户画像，需要处理海量的日志，花费大量时间和人力。尽管是如此高成本的事情，大部分公司还是希望能给自己的用户做一份足够精准的用户画像。

那么用户画像有什么作用，能帮助我们达到哪些目标呢？

大体上可以总结为以下几个方面：

1. 精准营销：精准直邮、短信、 App 消息推送、个性化广告等。
2. 用户研究：指导产品优化，甚至做到产品功能的私人定制等。
3. 个性服务：个性化推荐、个性化搜索等。
4. 业务决策：排名统计、地域分析、行业趋势、竞品分析等。

用户画像的内容

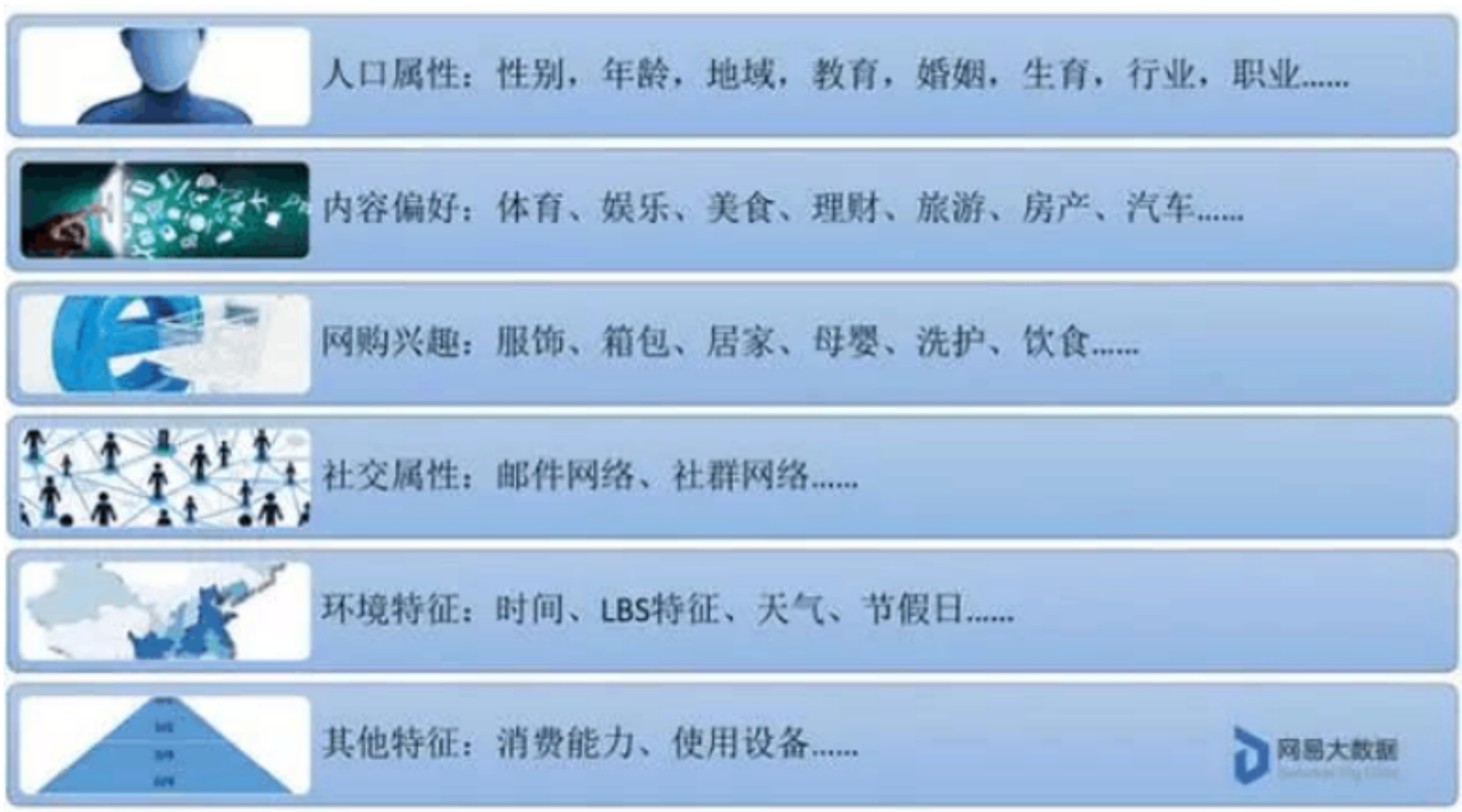
用户画像包含的内容并不完全固定，根据行业和产品的不同所关注的特征也有不同。对于大部分互联网公司，用户画像都会包含人口属性和行为特征。人口属性主要指用户的年龄、性别、所在的省份和城市、教育程度、婚姻情况、生育情况、工作所在的行业和职业等。行为特征主要包含活跃度、忠诚度等指标。

除了以上较通用的特征，不同类型的网站提取的用户画像各有侧重点。

- 以内容为主的媒体或阅读类网站，还有搜索引擎或通用导航类网站，往往会提取用户对浏览内容的兴趣特征，比如体育类、娱乐类、美食类、理财类、旅游类、房产类、汽车类等等。
- 社交网站的用户画像，也会提取用户的社交网络，从中可以发现关系紧密的用户群和在社群中起到意见领袖作用的明星节点。
- 电商购物网站的用户画像，一般会提取用户的网购兴趣和消费能力等指标。网购兴趣主要指用户在网购时的类目偏好，比如服饰类、箱包类、居家类、母婴类、洗护类、饮食类等。
- 消费能力指用户的购买力，如果做得足够细致，可以把用户的实际消费水平和在每个类目的心理消费水平区分开，分别建立特征纬度。

另外还可以加上用户的环境属性，比如当前时间、访问地点 LBS特征、当地天气、节假日情况等。

当然，对于特定的网站或 App，肯定又有特殊关注的用户纬度，就需要把这些维度做到更加细化，从而能给用户提供更精准的个性化服务和内容。

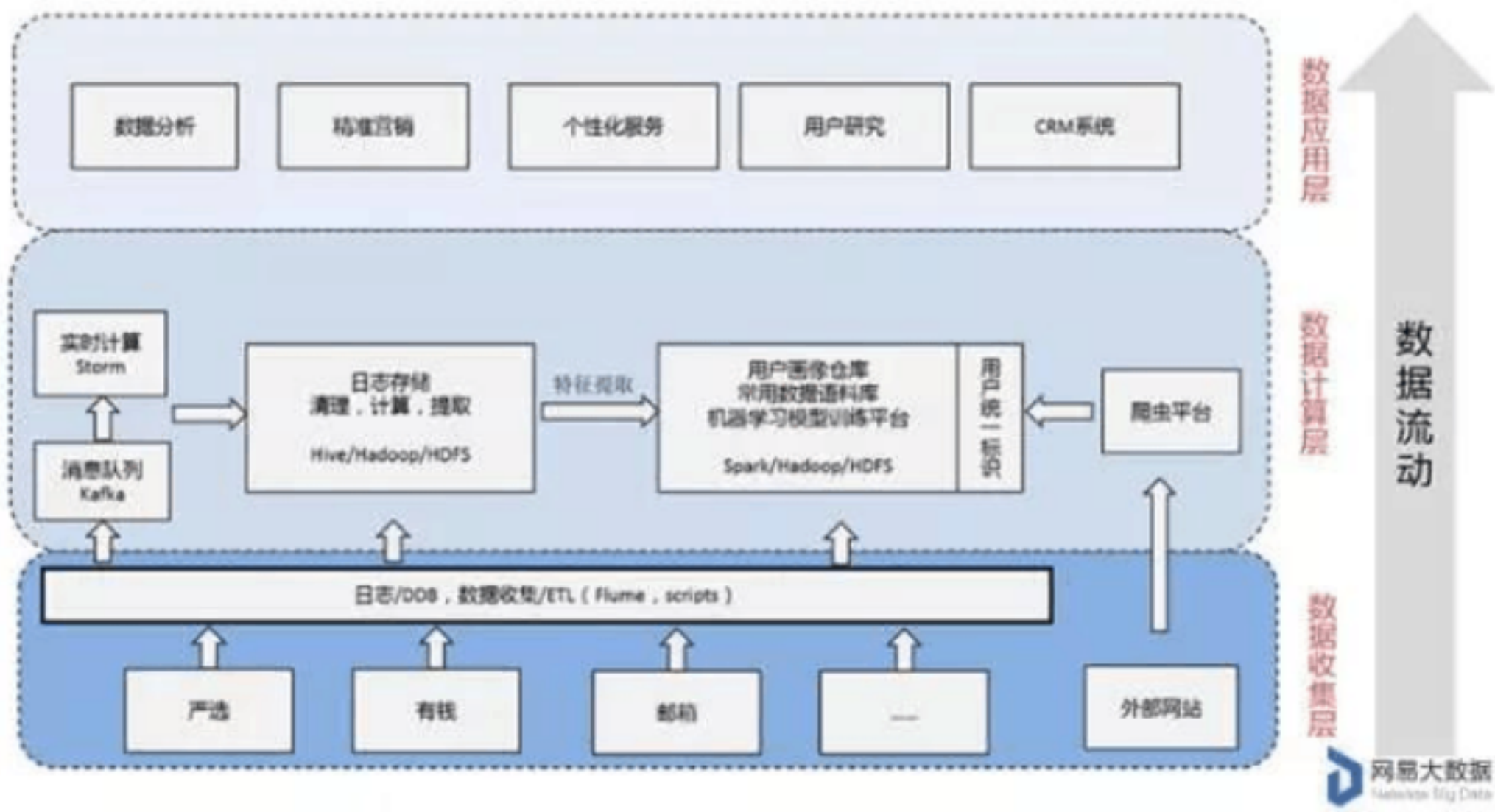


用户画像的生产

用户特征的提取即用户画像的生产过程，大致可以分为以下几步：

1. 用户建模，指确定提取的用户特征维度，和需要使用到的数据源。
2. 数据收集，通过数据收集工具，如 Flume 或自己写的脚本程序，把需要使用的数据统一存放到 Hadoop 集群。

3. 数据清理，数据清理的过程通常位于 Hadoop 集群，也有可能与数据收集同时进行，这一步的主要工作，是把收集到各种来源、杂乱无章的数据进行字段提取，得到关注的目标特征。
4. 模型训练，有些特征可能无法直接从数据清理得到，比如用户感兴趣的内容或用户的消费水平，那么可以通过收集到的已知特征进行学习和预测。
5. 属性预测，利用训练得到的模型和用户的已知特征，预测用户的未知特征。
6. 数据合并，把用户通过各种数据源提取的特征进行合并，并给出一定的可信度。
7. 数据分发，对于合并后的结果数据，分发到精准营销、个性化推荐、 CRM等各个平台，提供数据支持。



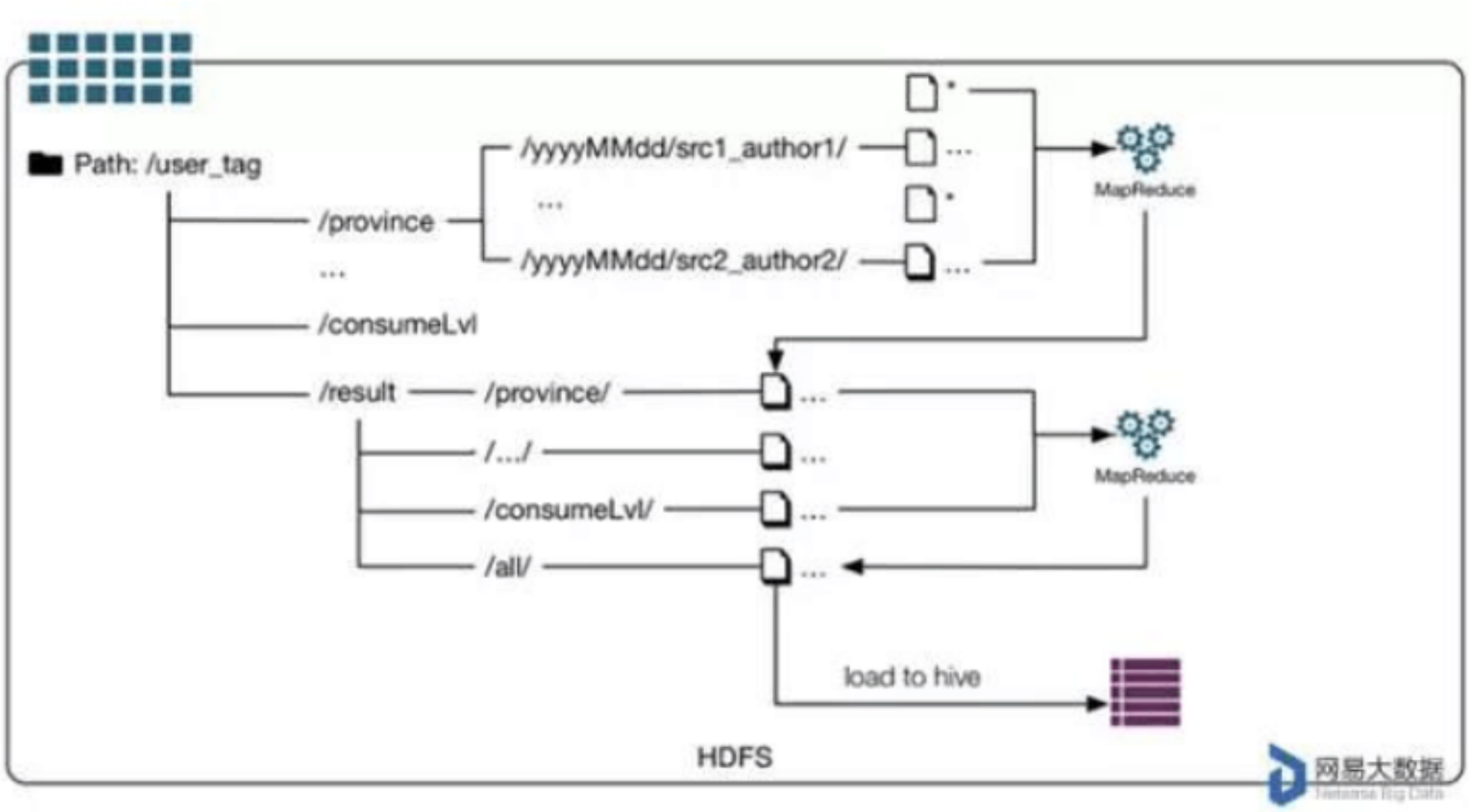
下面以用户性别为例，具体介绍特征提取的过程：

1. 提取用户自己填写的资料，比如注册时或者活动中填写的性别资料，这些数据准确率一般很高。
2. 提取用户的称谓，如文本中有提到的对方称呼，例如： xxx先生/女士，这个数据也比较准。
3. 根据用户姓名预测用户性别，这是一个二分类问题，可以提取用户的名字部分（百家姓与性别没有相关性），然后用朴素贝叶斯分类器训练一个分类器。过程中遇到了生僻字问题，比如 “甄嬛”的“嬛”，由于在名字中出现的少，因此分类器无法进行正确分类。考虑到汉字都是由偏旁部首组成，且偏旁部首也常常具有特殊含义（很多与性别具有相关性，比如草字头倾向女性，金字旁倾向男性），我们利用五笔输入法分解单字，再把名字本身和五笔打法的字母一起放到 LR分类器进行训练。比如，“嬛”字的打法：『 女 V+ 罒 L+ 一 G+ 衣 E = VLGE 』，这里的女字旁就很有女性倾向。
4. 另外还有一些特征可以利用，比如用户访问过的网站，经常访问一些美妆或女性服饰类网站，是女性的可能性就高；访问体育军事类网站，是男性的可能性就高。还有用户上网的时间段，经常深夜上网的用户男性的可能性就高。把这些特征加入到 LR分类器进行训练，也能提高一定的数据覆盖率。

数据管理系统

用户画像涉及到大量的数据处理和特征提取工作，往往需要用到多数据来源，且多人并行处理数据和生成特征。因此，需要一个数据管理系统来对数据统一进行合并存储和分发。我们的系统以约定的目录结构来组织数据，基本目录层级为： /user_tag/ 属性 / 日期 / 来源 _作者 /。以性别特征为例，开发者 dev1 从用户姓名提取的性别数据存放路径为 /user_tag/gender/20170101/name_dev1 ，开发者 dev2 从用户填写资料提取的性别数据存放路径为 /user_tag/gender/20170102/raw_dev2 。

从每种来源提取的数据可信度是不同的，所以各来源提取的数据必须给出一定的权重，约定一般为 0-1 之间的一个概率值，这样系统在做数据的自动合并时，只需要做简单的加权求和，并归一化输出到集群，存储到事先定义好的 Hive 表。接下来就是数据增量更新到 HBase、ES、 Spark集群等更多应用服务集群。



应用示例：个性化推荐

以电商网站的某种页面的个性化推荐为例，考虑到特征的可解释性、易扩展和模型的计算性能，很多线上推荐系统采用 LR（逻辑回归）模型训练，这里也以 LR模型举例。很多推荐场景都会用到基于商品的协同过滤，而基于商品协同过滤的核心是一个商品相关性矩阵 W，假设有 n个商品，那么 W就是一个 n * n 的矩阵，矩阵的元素 wij 代表商品 Ii和 Ij之间的相关系数。而根据用户访问和购买商品的行为特征，可以把用户表示成一个 n维的特征向量 U=[i1, i2, ..., in] 。于是 U*W 可以看成用户对每个商品的感兴趣程度 V=[v1, v2, ..., vn] ，这里 v1即是用户对商品 I1的感兴趣程度， v1= i1*w11 + i2*w12 + in*w1n 。如果把相关系数 w11, w12, ..., w1n 看成要求的变量，那么就可以用 LR模型，代入训练集用户的行为向量 U，进行求解。这样一个初步的 LR模型就训练出来了，效果和基于商品的协同过滤类似。 这时只用到了用户的行为特征部分，而人口属性、网购偏好、内容偏好、消费能力和环境特征等其他上下文还没有利用起来。把以上特征加入到 LR模型，同时再加上目标商品自身的属性，如文本标签、所属类目、销量等数据，如下图所示，进一步优化训练原来的 LR模型。从而最大程度利用已经提取的用户画像数据，做到更精准的个性化推荐。

点评

用户画像是当前大数据领域的一种典型应用，也普遍应用在多款网易互联网产品中。本文基于网易的实践，深入浅出地解析了用户画像的原理和生产流程。

精确有效的用户画像，依赖于从大量的数据中提取正确的特征，这需要一个强大的数据管理系统作为支撑。网易大数据产品体系中包含的一站式大数据开发与管理平台 –网易猛犸，正是在网易内部实践中打磨形成的，能

够为用户画像及后续的业务目标实现提供数据传输、计算和作业流调度等基础能力，有效降低大数据应用的技术门槛。